## **Part 4: Short-Answer Questions (Upload to Blackboard)**

Answer the following based on your work in **Part 3**:

1. Please provide the link to your public **GitHub repository**.

https://github.com/Binal-1805/BINF-5507

**Regression Models:**

1. Explain how ElasticNet regularization balances L1 and L2 penalties. Why might this be advantageous for certain datasets?

Ans: ElasticNet regularization combines both L1 and L2 penalties, allowing it to balance the strengths of both.

L1- Penalty encourages sparsity by setting some coefficients exactly to zero, which helps feature selection

L2- Penalty shrinks the coefficients smoothly, preventing them from becoming too large.

Advantageous when:

* + - 1. When the dataset has many correlated features (where L2 helps by distributing the penalty across them)
      2. When some features should be entirely exculded (where L1’s sparsity is useful)

1. What does the heatmap of R² and RMSE reveal about the relationship between alpha, l1\_ratio, and model performance? How did you determine the optimal configuration?

Ans: The heatmap of R² and RMSE shows how alpha and l1\_ratio impact model performance in ElasticNet. A high alpha value typically leads to stronger regularization, reducing overfitting but possibly lowering model performance. The l1\_ratio influences the balance between Lasso (L1) and Ridge (L2) regularization, affecting both sparsity and model complexity. To determine the optimal configuration, I examined the heatmap for the highest R² and the lowest RMSE. This allowed me to select the combination of alpha and l1\_ratio that maximized model accuracy while preventing overfitting.

**Classification Models:**

1. When comparing logistic regression and k-NN, what evaluation metric(s) did you prioritize, and why?

Ans: When comparing logistic regression and k-NN, I prioritized the following evaluation metrics:

* + - 1. Accuracy
      2. F1 Score:
      3. AUROC (Area Under ROC Curve
      4. **AUPRC (Area Under Precision-Recall Curve**

**Prioritize these matrics because:**

**In medical datasets like heart disease, where class imbalance is common, accuracy can be misleading. Metrics like F1 Score, AUROC, and AUPRC are more useful as they better evaluate model performance, especially for the minority class. AUROC and AUPRC offer insights into model robustness across different decision thresholds, making them more reliable than accuracy alone.**

This metrics together offer a comprehensive view of how well each model (Logistic Regression and k-NN) performs in distinguishing between heart disease presence and absence while accounting for class imbalance.

1. What insights did you gain from comparing AUROC and AUPRC curves for the top-performing models? Which model would you recommend, and under what circumstances?

Ans: AUROC evaluates overall model performance, while AUPRC is more reliable for imbalanced datasets, focusing on the minority class. For imbalanced medical datasets like heart disease, AUPRC is preferred. I’d recommend the model with the best balance of precision and recall, prioritizing higher recall if false negatives are costly.

#### **Critical Thinking (BONUS)**

1. How do different solvers (e.g., liblinear, saga) affect the behavior of logistic regression models? Which solver worked best in your experiments and why?
2. For k-NN, what trade-offs arise from increasing or decreasing the value of n\_neighbors, and how does this impact model complexity?